A MEASURE OF DISPERSION FOR ORDERED SERIES*

BY W. L. CRUM, *Williamstown, Massachusetts*

It is the object of this paper to call attention to the inadequacy of the standard deviation for the study of the dispersion of a statistical series the terms of which are ordered relative to a given variable, to examine certain considerations bearing upon the dispersion in such series, and to set up tentatively a new measure particularly applicable to ordered series.

I. EXISTING MEASURES

A statistical series may be assigned to one of two broad classes according as it consists merely of a list of numbers of indefinite arrangement, or has its items ordered relative to a particular variable. Typical of the first class are the series composed of experimental measurements, and chief illustrations of the second class are to be found in historical series in the field of economic statistics. The common statistical coefficients have been developed in the study of, and are peculiarly applicable to, series of the first class; and it is a question of considerable moment whether such coefficients are equally useful in the analysis of an ordered series of the second class.

The chief measure of dispersion is the standard deviation, the square root of the mean squared deviation from the arithmetic mean. It is apparent from the method of calculation of the standard deviation that it can take no account of the arrangement of the variates.

Example (i). Consider the two series

(a) 3 7 4 6 2 5 8 7 3 4 7 3 5 8 2 6

(b) 2 3 2 3 4 5 5 6 7 7 8 7 8 6 3 4

obtained by arranging in two different orders the items

(c) 2 2 3 3 3 4 4 5 5 6 6 7 7 7 8 8.

The standard deviation for both (a) and (b) is the same as for (c), namely, 1.94. Nevertheless, inspection of the series (a) and (b), or a glance at their representation in Figure 1, is sufficient to reveal the striking difference between the two: (a) fluctuates violently, whereas (b) advances with fairly stable sweep. This coefficient fails, therefore, to distinguish the erratic series (a) from the fairly smooth (b).

* Read before the American Mathematical Society on September 9, 1921, under the title "A tentative substitute for the standard deviation in the examination of the dispersion of an ordered statistical series."

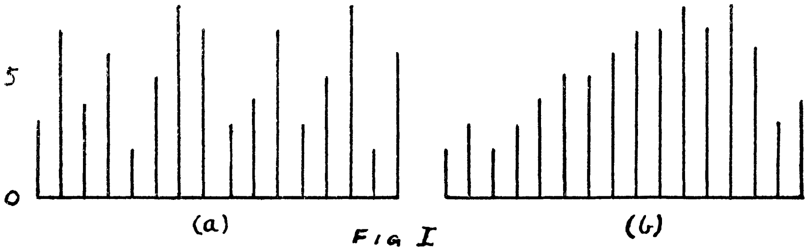


FIG I

Similar objections may be raised to the other common measures of dispersion. The coefficient of variation is derived from the standard deviation and has the same shortcoming. The computation of the mean deviation is independent of the order of arrangement of the variates, and hence it can give no better result than the standard deviation. The quartile deviation involves a grouping of the variates according to their size, regardless of the actual arrangement of the variates in their original form.

The conclusion seems warranted that existing measures of dispersion are not fitted to distinguish between such series as those of example (i). These coefficients attach the same importance to a given difference between two variates which are widely separated and two which are adjacent. They indicate the extent of the fluctuation, but fail to account for its rate. For certain purposes in the study of ordered series, particularly historical economic series, it would seem to be this very rate which it is of importance to examine and measure. Since the term *dispersion* is so widely understood in its present sense and is so generally accepted as the quality measured by the standard deviation, it may be well to avoid speaking of this rate of fluctuation as dispersion. We adopt provisionally the term *fluctuation-rate.*

II. CORRELATION COEFFICIENT

We seek now to measure the fluctuation-rate of an order series. It is suggested elsewhere* that for an ordered series possessing rectilinear trend, a more reliable measure of fluctuation than the ordinary standard deviation, σ_x, is the generalized standard deviation

$$\sigma_{x.t} = \sigma_x \sqrt{1 - r^2_{xt}}$$

since this latter in effect eliminates that part of the fluctuation due to the trend. Now the quantity r_{xt} in the above formula is the coefficient of correlation between the x_i—the series being examined—and

* Crum, W. L., *The Significance of the Partial Correlation Coefficient in the Comparison of Ordered Statistical Series Possessing Rectilinear Trends.* (Quarterly Publications of the American Statistical Association, Vol. XVII, pp. 949–952.)

the t_i—the variable relative to which the x_i are ordered. We examine in detail below the sufficiency of r_{xt} as a measure of fluctuation-rate, and any objections found will hold also against $\sigma_{x.t}$.

It would seem at first glance that the correlation coefficient r_{xt} might serve as a measure of fluctuation-rate: the smaller its value, the nearer a straight line is the join of the tops of the ordinates representing the variates, and the less is the fluctuation. It is at once necessary, however, to inquire whether a series of variates does not admit of several arrangements, differing markedly in rate of fluctuation, but all having the same value for the coefficient of correlation of the x_i relative to the t_i.

In fact, if there are N variates,

$$r_{xt}=\frac{\Sigma(t_i-T)(x_i-X)}{N\sigma_x\sigma_t}$$

where T and X are the arithmetic means and σ_t and σ_x are the standard deviations of the t_i and x_i respectively. It is evident that for a given group of x_i, N, σ_t, and σ_x will all be fixed constants, regardless of the order of arrangement of the x_i relative to the t_i. The value of r_{xt} for the various arrangements will therefore depend upon the product sum

$$Np=\Sigma(t_i-T)(x_i-X).$$

Moreover, it is easily shown that this value of Np differs from

$$A=\Sigma t_i x_i$$

only by the constant, NXT. Therefore, it is sufficient to examine the various values of A which correspond to the different arrangements in order to test the adequacy of r_{xt} as a measure of fluctuation-rate.

The question now before us is: Are there several arrangements of a given group of x_i relative to the t_i, such that for all these arrangements A is identical? If it could be shown that there is only one arrangement corresponding to a given value of A, r_{xt} would indeed be an ideal index of fluctuation-rate. Or, if it develops that all the possible arrangements giving a particular value of A are such that they do not differ from one another in the character sought to be measured—the fluctuation-rate—r_{xt} would be entirely adequate to the purpose. Unfortunately, it proves that r_{xt} falls short of meeting even this second requirement.

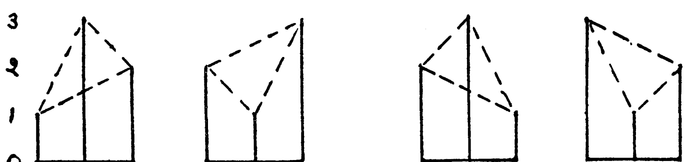
Without attempting at present an analytical study of the above question, we can by experiment with numerical examples bring out the essential fact about the failure of r_{xt} to discriminate properly between the various arrangements.

Example (ii). Suppose the x_i consist of the intergers 1, 2, 3, to be arranged arbitrarily relative to $t_i = 1, 2, 3$. The possible arrangements are:

		(a)	(b)	(c)	(d)	(e)	(f)
t_i	A	14	13	11	13	11	10
1	x_1	1	1	2	2	3	3
2	x_2	2	3	3	1	1	2
3	x_3	3	2	1	3	2	1

the value of A being given at the head of each series.

Series (b) and (d) each have A 13, and they are clearly similar in their fluctuation-rates. This is evident from the series themselves, or from Figure 2. Indeed, the nature of the fluctuation may be inferred from the two dotted triangles (this aspect of the question will be viewed more fully below), and the one triangle is obviously obtainable from the other by a succession of reflections and inversions. Similar remarks apply to the two series (c) and (e), for which A is 11.



(b) (d) (c) (e)

FIG. II

On the other hand, although A is 13 for (b) and 11 for (c), these two series have identical fluctuation-rates; for (c) is merely (b) in reverse order. The same is true of series (d) and (e), and the objection is even clearer in the case of (a) and (f): these pairs each have a single fluctuation-rate, but the value of A differs for the two members of the pair. It is apparent, then, that A does not have the same value for all arrangements having the same fluctuation-rate; but this is not the real question, which is whether all arrangements having the same A have equal fluctuation-rates.

Example (iii). Let the x_i be the integers from 1 to 4. The 24 possible arrangements can be classified in the following frequency table according to the value of A in the several cases:

A	20	21	22	23	24	25	26	27	28	29	30
freq.	1	3	1	4	2	2	2	4	1	3	1.

It is found, on inspection, that the series having A 20 and 30 are reverses of each other, and similarly those having A 21 and 29 are re-

verses in pairs, and so on throughout. Hence, the objection found under example (ii) arises here also, and in even more striking degree.

We examine next in detail the four series having A 27. They are:

(a) 1, 3, 4, 2; (b) 1, 4, 2, 3; (c) 2, 3, 1, 4; (d) 3, 1, 2, 4

and their diagrams appear in Figure 3. The series themselves and the diagrams show that (b) and (c) constitute a pair having equal fluctuation-rates, and so do (a) and (d); but the two pairs differ distinctly from each other. Of course, since they are the reverses of the above four series, the four series having A 23 are also grouped in pairs. Moreover, if we test the three series having A 29, it will be found that only two of them have identical fluctuation-rates.

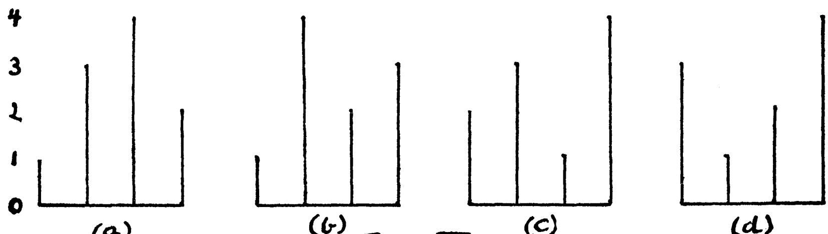


FIG III

It appears, therefore, that even for series of the simple sort given in this last example, the value of A—and hence of r_{xt}—does not furnish an adequate measure of the fluctuation-rate: it varies when the fluctuation-rate remains unchanged, and changes when the fluctuation-rate is constant. As the series become more complicated we shall have diminishing confidence in the sufficiency of the correlation coefficient for this measurement. We must hesitate even to say that it is surely better than the standard deviation; for whereas the latter gives one value for all 24 cases in problem (iii), the correlation coefficient does indeed make a distinction; but we have seen that it makes a distinction where none exists in the fluctuation-rate.

III. THE MEAN SQUARED SECOND ORDER DIFFERENCE

We recall that one way of recognizing the equality of the fluctuation-rates of two series was by the geometric correspondence of the triangles joining the tops of the ordinates. This was most apparent in the simplest case, shown in example (ii) and Figure 2. The area of the dotted triangle suggests itself as a measure of the fluctuation-rate. In a series of more than three variates there would be a series of these triangles, each belonging to three adjacent variates: for N variates

there would be $N-2$ triangles. At any point of the ordered series the fluctuation-rate might be measured by the area of the triangle belonging to the three nearest variates; and, for the whole series, the fluctuation-rate would be indicated by an average of the $N-2$ triangular areas.

If we assume for simplicity that the t_i are integers, the area of the triangle belonging to three successive variates x_{i-1}, x_i, x_{i+1}, is

$$\text{Area} = \tfrac{1}{2}(x_{i-1}+x_{i+1})-x_i.$$

For series (b) and (d) of example (ii) we get for the triangular area $-3/2$ and $3/2$ respectively; and these are equal if the sign is neglected. This is in accord with our inferences from the series and the figures.

In averaging the $N-2$ triangles in a series of N x_i, we may neglect the signs and take the simple arithmetic mean, or we may take the square root of the mean squared area. We choose the latter, and have for the average area:

$$\sqrt{\frac{1}{N-2}\sum_{2}^{N-1}\left(\frac{x_{i-1}+x_{i+1}}{2}-x_i\right)^2}.$$

The type expression in this summation is the square of

$$\tfrac{1}{2}(x_{i-1}+x_{i+1})-x_i$$

which is

$$\tfrac{1}{2}(x_{i+1}-2x_i+x_{i-1}).$$

This is precisely $\frac{1}{2}$ the second order finite difference, Δ''_i. Hence the average triangular area is one-half of

$$F=\sqrt{\frac{1}{N-2}\sum_{2}^{N-1}\Delta''^2_i}$$

and, if we discard the $\frac{1}{2}$, we may take F as the new coefficient to measure the fluctuation-rate, viz., the square root of the mean squared second order difference.

We arrive at F as a measure of the fluctuation-rate by study of the deviation triangles, but the present definition in terms of the second differences gives added reason for accepting it. In fact, the first order finite difference Δ'_i is concerned with the slope of the join of two successive ordinates, whereas the second order difference Δ''_i has to do with the curvature of the join of three successive ordinates. It is this curvature, averaged over the entire extent of the series, that it is sought to measure.

It is clear that the actual calculation of F is very simple, and indeed scarcely more complicated than that of σ. Although it is necessary to calculate both the first and second differences, it is not necessary to make any correction for the position of the mean. As an example, we calculate F for the two series of example (i).

Case (a)

x_i 3 7 4 6 2 5 8 7 3 4 7 3 5 8 2 6

Δ'_i 4 −3 2 −4 3 3 −1 −4 1 3 −4 2 3 −6 4

Δ''_i −7 5 −6 7 0 −4 −3 5 2 −7 6 1 −9 10

$F = 5.85$

Case (b)

x_i 2 3 2 3 4 5 5 6 7 7 8 7 8 6 3 4

Δ'_i 1 −1 1 1 1 0 1 1 0 1 −1 1 −2 −3 1

Δ''_i −2 2 0 0 −1 1 0 −1 1 −2 2 −3 −1 4

$F = 1.81$

Evidently F distinguishes readily the erratic series (a) from the fairly regular (b).

We turn now to those series of example (iii) for which A is 27. The values of F for the four series are:

(a) 2.24; (b) 4.12; (c) 4.12; (d) 2.24.

We recall from Figure 3 that (c) is the "reflection" of (b), and has the same fluctuation-rate. This fact, which was not, to be sure, out of accord with the single value 27 for A, is borne out by this new coefficient. The same is true of (a) and (d). On the other hand, the distinction between the two pairs, which was not indicated by any change in A, is clearly shown by the differing value of F. Moreover, a calculation of the values of F for the four series having A 23 shows the same values as above, and the series in the A 23 group which are the reverses of (a) and (d) in the A 27 group have also F equal to 2.24, and similarly for the other pair. Similar tests apply throughout, and we may, therefore, say that both the objections raised against A are met by F.

Although we are not yet in a position to say with finality that F furnishes the best measure of the fluctuation-rate in an ordered statistical series, the tests which we have made above seem to indicate that it may be accepted tentatively as such measure. It is hoped that further study of the coefficient F will serve to develop more fully the real significance of the square root of the mean squared second order difference, and perhaps result in devising another measure even more efficacious than F.